1

Introduction to Ridge Regression

This chapter reviews the developments of ridge regression, starting with the definition of ridge regression together with the covariance matrix. We discuss the multicollinearity problem and ridge notion and present the preliminary test and Stein-type estimators. In addition, we discuss the high-dimensional problem. In conclusion, we include detailed notes, references, and organization of the book.

1.1 Introduction

Consider the common multiple linear regression model with the vector of coefficients, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ given by

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1.1}$$

where $\boldsymbol{Y} = (y_1, \ldots, y_n)^\top$ is a vector of n responses, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$ is an $n \times p$ design matrix of rank p ($\leq n$), $\boldsymbol{x}_i \in \mathbb{R}^p$ is the vector of covariates, and $\boldsymbol{\epsilon}$ is an n-vector of independently and identically distributed (i.i.d.) random variables (r.v.).

The least squares estimator (LSE) of $\boldsymbol{\beta}$, denoted by $\tilde{\boldsymbol{\beta}}_n$, can be obtained by minimizing the residual sum of squares (RSS), the convex optimization problem,

$$\min_{\boldsymbol{\beta}}\{(Y - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})\} = \min_{\boldsymbol{\beta}}\{S(\boldsymbol{\beta})\},$$

where $S(\boldsymbol{\beta}) = \boldsymbol{Y}^\top\boldsymbol{Y} - 2\boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{Y} + \boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta}$ is the RSS. Solving

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}^\top\boldsymbol{Y} + 2\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0}$$

Theory of Ridge Regression Estimation with Applications, First Edition.
A.K. Md. Ehsanes Saleh, Mohammad Arashi, and B.M. Golam Kibria.

with respect to (w.r.t.) $\boldsymbol{\beta}$ gives

$$\tilde{\boldsymbol{\beta}}_n = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}. \tag{1.2}$$

Suppose that $\mathbb{E}(\boldsymbol{\epsilon}) = 0$ and $\mathbb{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top) = \sigma^2 \boldsymbol{I}_n$ for some $\sigma^2 \in \mathbb{R}^+$. Then, the variance–covariance matrix of LSE is given by

$$\mathrm{Var}(\tilde{\boldsymbol{\beta}}_n) = \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}. \tag{1.3}$$

Now, we consider the canonical form of the multiple linear regression model to illustrate how large eigenvalues of the design matrix $\boldsymbol{X}^\top \boldsymbol{X}$ may affect the efficiency of estimation.

Write the spectral decomposition of the positive definite design matrix $\boldsymbol{X}^\top \boldsymbol{X}$ to get $\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{\Gamma} \boldsymbol{\Lambda} \boldsymbol{\Gamma}^\top$, where $\boldsymbol{\Gamma}(p \times p)$ is a column orthogonal matrix of eigenvectors and $\boldsymbol{\Lambda} = \mathrm{Diag}(\lambda_1, \ldots, \lambda_p)$, where $\lambda_j > 0$, $j = 1, \ldots, p$ is the ordered eigenvalue matrix corresponding to $\boldsymbol{X}^\top \boldsymbol{X}$. Then,

$$\boldsymbol{Y} = \boldsymbol{T}\boldsymbol{\xi} + \boldsymbol{\epsilon}, \quad \boldsymbol{T} = \boldsymbol{X}\boldsymbol{\Gamma}, \quad \boldsymbol{\xi} = \boldsymbol{\Gamma}^\top \boldsymbol{\beta}. \tag{1.4}$$

The LSE of $\boldsymbol{\xi}$ has the form,

$$\begin{aligned} \tilde{\boldsymbol{\xi}}_n &= (\boldsymbol{T}^\top \boldsymbol{T})^{-1} \boldsymbol{T}^\top \boldsymbol{Y} \\ &= \boldsymbol{\Lambda}^{-1} \boldsymbol{T}^\top \boldsymbol{Y}. \end{aligned} \tag{1.5}$$

The variance–covariance matrix of $\tilde{\boldsymbol{\xi}}_n$ is given by

$$\mathrm{Var}(\tilde{\boldsymbol{\xi}}_n) = \sigma^2 \boldsymbol{\Lambda}^{-1} = \sigma^2 \begin{bmatrix} \frac{1}{\lambda_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\lambda_2} & 0 & 0 \\ \vdots & \cdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\lambda_p} \end{bmatrix}. \tag{1.6}$$

Summation of the diagonal elements of the variance–covariance matrix of $\tilde{\boldsymbol{\xi}}_n$ is equal to $\mathrm{tr}(\mathrm{Var}(\tilde{\boldsymbol{\xi}}_n)) = \sigma^2 \sum_{j=1}^{p} \lambda_j^{-1}$. Apparently, small eigenvalues inflate the total variance of estimate or energy of $\boldsymbol{X}^\top \boldsymbol{X}$. Specifically, since the eigenvalues are ordered, if the first eigenvalue is small, it causes the variance to explode. If this happens, what must one do? In the following section, we consider this problem. Therefore, it is of interest to realize when the eigenvalues become small.

Before discussing this problem, a very primitive understanding is that if we enlarge the eigenvalues from λ_j to $\lambda_j + k$, for some positive value, say, k, then we can prevent the total variance from exploding. Of course, the amount of recovery depends on the correct choice of the parameter, k.

An artificial remedy is to have $\mathrm{tr}(\mathrm{Var}(\tilde{\boldsymbol{\xi}}_n)) = \sigma^2 \sum_{j=1}^{p} (\lambda_j + k)^{-1}$ based on the variance matrix given by

$$\mathrm{Var}(\tilde{\boldsymbol{\xi}}_n) = \sigma^2 \begin{pmatrix} \frac{1}{\lambda_1 + k} & 0 & \cdots & 0 \\ 0 & \frac{1}{\lambda_2 + k} & 0 & 0 \\ \vdots & \cdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\lambda_p + k} \end{pmatrix} = \sigma^2 (\boldsymbol{\Lambda} + k\boldsymbol{I}_p)^{-1}, \quad k \in \mathbb{R}^+. \tag{1.7}$$

Replacing the eigenvector matrix $\boldsymbol{\Gamma}$ in (1.5) by this matrix (1.7), we get the $\tilde{\boldsymbol{\xi}}_n = (\boldsymbol{\Lambda} + k\boldsymbol{I}_p)^{-1}\boldsymbol{T}^\top \boldsymbol{Y}$ and the variance as in (1.8).

$$\begin{aligned} \mathrm{Var}(\tilde{\boldsymbol{\xi}}_n) &= (\boldsymbol{\Lambda} + k\boldsymbol{I}_p)^{-1}\boldsymbol{T}^\top \mathrm{Var}(\boldsymbol{Y})\boldsymbol{T}(\boldsymbol{\Lambda} + k\boldsymbol{I}_p)^{-1} \\ &= \sigma^2 (\boldsymbol{\Lambda} + k\boldsymbol{I}_p)^{-1}\boldsymbol{\Lambda}(\boldsymbol{\Lambda} + k\boldsymbol{I}_p)^{-1}, \end{aligned} \tag{1.8}$$

which shows

$$\begin{aligned} \mathrm{tr}(\mathrm{Var}(\tilde{\boldsymbol{\xi}}_n)) &= \sigma^2 \mathrm{tr}((\boldsymbol{\Lambda} + k\boldsymbol{I}_p)^{-1}\boldsymbol{\Lambda}(\boldsymbol{\Lambda} + k\boldsymbol{I}_p)^{-1}) \\ &= \sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k)^2} \end{aligned} \tag{1.9}$$

Further, we show that achieving the total variance of $\sigma^2 \sum_{j=1}^{p} \lambda_j/(\lambda_j + k)^2$ is the target.

1.1.1 Multicollinearity Problem

Multicollinearity or collinearity is the existence of near-linear relationships among the regressors, predictors, or input/exogenous variables. There are terms such as exact, complete and severe, or supercollinearity and moderate collinearity. Supercollinearity indicates that two (or multiple) covariates are linearly dependent, and moderate occurs when covariates are moderately correlated. In the complete collinearity case, the design matrix is not invertible. This case mostly occurs in a high-dimensional situation (e.g. microarray measure) in which the number of covariates (p) exceeds the number of samples (n).

Moderation occurs when the relationship between two variables depends on a third variable, namely, the moderator. This case mostly happens in structural equation modeling. Although moderate multicollinearity does not cause the mathematical problems of complete multicollinearity, it does affect the

interpretation of model parameter estimates. According to Montgomery et al. (2012), if there is no linear relationship between the regressors, they are said to be orthogonal.

Multicollinearity or ill-conditioning can create inaccurate estimates of the regression coefficients, inflate the standard errors of the regression coefficients, deflate the partial t-tests for the regression coefficients, give false and nonsignificant p-values, and degrade the predictability of the model. It also causes changes in the direction of signs of the coefficient estimates. According to Montgomery et al. (2012), there are five sources for multicollinearity: (i) data collection, (ii) physical constraints, (iii) overdefined model, (iv) model choice or specification, and (v) outliers.

There are many studies that well explain the problem of multicollinearity. Since theoretical aspects of ridge regression and related issues are our goal, we refer the reader to Montgomery et al. (2012) for illustrative examples and comprehensive study on the multicollinearity and diagnostic measures such as correlation matrix, eigen system analysis of $\boldsymbol{X}^\top\boldsymbol{X}$, known as condition number, or variance decomposition proportion and variance inflation factor (VIF). To end this section, we consider a frequently used example in a ridge regression, namely, the Portland cement data introduced by Woods et al. (1932) from Najarian et al. (2013). This data set has been analyzed by many authors, e.g. Kaciranlar et al. (1999), Kibria (2003), and Arashi et al. (2015). We assemble the data as follows:

$$\boldsymbol{X} = \begin{bmatrix} 7 & 26 & 6 & 60 \\ 1 & 29 & 15 & 52 \\ 11 & 56 & 8 & 20 \\ 11 & 31 & 8 & 47 \\ 7 & 52 & 6 & 33 \\ 11 & 55 & 9 & 22 \\ 3 & 71 & 17 & 6 \\ 1 & 31 & 22 & 44 \\ 2 & 54 & 18 & 22 \\ 21 & 47 & 4 & 26 \\ 1 & 40 & 23 & 34 \\ 11 & 66 & 9 & 12 \\ 10 & 68 & 8 & 12 \end{bmatrix}, \quad \boldsymbol{Y} = \begin{bmatrix} 78.5 \\ 74.3 \\ 104.3 \\ 87.6 \\ 95.9 \\ 109.2 \\ 102.7 \\ 72.5 \\ 93.1 \\ 115.9 \\ 83.8 \\ 113.3 \\ 109.4 \end{bmatrix}. \tag{1.10}$$

The following (see Tables 1.1 and 1.2) is the partial output of linear regression fit to this data using the software SPSS, where we selected Enter as the method.

We display the VIF values to diagnose the multicollinearity problem in this data set. Values greater than 10 are a sign of multicollinearity. Many softwares can be used.

Table 1.1 Model fit indices for Portland cement data.

Model	R	R^2	Adjacent R^2	Standard error of the estimate
1	0.991	0.982	0.974	2.446

Table 1.2 Coefficient estimates for Portland cement data.

Coefficient	Standardized coefficient	t	Significance	VIF
(Constant)		0.891	0.399	
x_1	0.607	2.083	0.071	38.496
x_2	0.528	0.705	0.501	254.423
x_3	0.043	0.135	0.896	46.868
x_4	−0.160	−0.203	0.844	282.513

1.2 Ridge Regression Estimator: Ridge Notion

If the regression coefficients β_js are unconstrained, then they can explode (become large); this results in high variance. Hence, in order to control the variance, one may regularize the regression coefficients and determine how large the coefficient grows. In other words, one may impose a constraint on them so as not to get unboundedly large or penalized large regression coefficients. One type of constraint is the ridge constraint given by $\sum_{j=1}^{p} \beta_j^2 \leq t$ for some positive value t. Hence, the minimization of the penalized residual sum of squares (PRSS) is equivalent to solving the following convex optimization problem,

$$\min_{\boldsymbol{\beta}} \{(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})\} \quad \text{such that} \quad \sum_{j=1}^{p} \beta_j^2 \leq t, \tag{1.11}$$

for some positive value t.

In general, the PRSS is defined by

$$(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + k\|\boldsymbol{\beta}\|^2, \quad \|\boldsymbol{\beta}\|^2 = \sum_{j=1}^{p} \beta_j^2. \tag{1.12}$$

Since the PRSS is a convex function w.r.t. $\boldsymbol{\beta}$, it has a unique solution. Because of the ridge constraint, the solution is termed as the ridge regression estimator (RRE).

To derive the RRE, we solve the following convex optimization problem

$$\min_{\boldsymbol{\beta}}\{(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})+k\|\boldsymbol{\beta}\|^2\}=\min_{\boldsymbol{\beta}}\{\mathrm{PS}(\boldsymbol{\beta})\}, \tag{1.13}$$

where $\mathrm{PS}(\boldsymbol{\beta})=\boldsymbol{Y}^\top\boldsymbol{Y}-2\boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{Y}+\boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta}+k\boldsymbol{\beta}^\top\boldsymbol{\beta}$ is the PRSS. Solving

$$\frac{\partial\mathrm{PS}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}=-2\boldsymbol{X}^\top\boldsymbol{Y}+2\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta}+2k\boldsymbol{\beta}=\mathbf{0}$$

w.r.t. $\boldsymbol{\beta}$ gives the RRE,

$$\hat{\boldsymbol{\beta}}_n^{\mathrm{RR}}(k)=(\boldsymbol{X}^\top\boldsymbol{X}+k\boldsymbol{I}_p)^{-1}\boldsymbol{X}^\top\boldsymbol{Y}. \tag{1.14}$$

Here, k is the shrinkage (tuning) parameter. Indeed, k tunes (controls) the size of the coefficients, and hence regularizes them. As $k\to 0$, the RRE simplifies to the LSE. Also, as $k\to\infty$, the RREs approach zero. Hence, the optimal shrinkage parameter k is of interest.

One must note that solving the optimization problem (1.13) is not the only way of yielding the RRE. It can also be obtained by solving a RSS of another data, say augmented data. To be specific, consider the following augmentation approach. Let

$$\boldsymbol{X}^*=\begin{bmatrix}\boldsymbol{X}\\ \sqrt{k}\boldsymbol{I}_p\end{bmatrix},\quad \boldsymbol{Y}^*=\begin{bmatrix}\boldsymbol{Y}\\ \mathbf{0}\end{bmatrix}. \tag{1.15}$$

Assume the following multiple linear model,

$$\boldsymbol{Y}^*=\boldsymbol{X}^*\boldsymbol{\beta}+\boldsymbol{\epsilon}^*, \tag{1.16}$$

where $\boldsymbol{\epsilon}^*$ is an $(n+p)$-vector of i.i.d. random variables. Then, the LSE of $\boldsymbol{\beta}$ is obtained as

$$\begin{aligned}\boldsymbol{\beta}_n^*&=\min_{\boldsymbol{\beta}}\{(\boldsymbol{Y}^*-\boldsymbol{X}^*\boldsymbol{\beta})^\top(\boldsymbol{Y}^*-\boldsymbol{X}^*\boldsymbol{\beta})\}\\ &=(\boldsymbol{X}^{*\top}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*\top}\boldsymbol{Y}^*\\ &=(\boldsymbol{X}^\top\boldsymbol{X}+k\boldsymbol{I}_p)^{-1}\boldsymbol{X}^\top\boldsymbol{Y}\\ &=\hat{\boldsymbol{\beta}}_n^{\mathrm{RR}}(k).\end{aligned} \tag{1.17}$$

Thus, the LSE of the augmented data is indeed the RRE of the normal data.

1.3 LSE vs. RRE

Indeed, the RRE is proportional to the LSE. Under the orthonormal case, i.e. $\boldsymbol{X}^\top\boldsymbol{X}=\boldsymbol{I}_p$, the RRE simplifies to

$$\begin{aligned}\hat{\boldsymbol{\beta}}_n^{\mathrm{RR}}(k)&=(\boldsymbol{X}^\top\boldsymbol{X}+k\boldsymbol{I}_p)^{-1}\boldsymbol{X}^\top\boldsymbol{Y}\\ &=(1+k)^{-1}\boldsymbol{I}_p\boldsymbol{X}^\top\boldsymbol{Y}\\ &=(1+k)^{-1}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y}\\ &=\frac{1}{1+k}\tilde{\boldsymbol{\beta}}_n.\end{aligned} \tag{1.18}$$

Theorem 1.1 *The ridge penalty shrinks the eigenvalues of the design matrix.*

Proof: Write the singular value decomposition (SVD) of the design matrix $\boldsymbol{X}$ to get $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^\top$, where $\boldsymbol{U}$ $(n \times p)$ and $\boldsymbol{V}$ $(p \times p)$ are column orthogonal matrices, and $\boldsymbol{\Lambda} = \text{Diag}(\lambda_1, \ldots, \lambda_p)$ $\lambda_j > 0, j = 1, \ldots, p$ is the eigenvalue matrix corresponding to $\boldsymbol{X}^\top\boldsymbol{X}$. Then,

$$\begin{aligned}
\tilde{\boldsymbol{\beta}}_n &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y} \\
&= (\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^\top)^{-1}\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{U}^\top\boldsymbol{Y} \\
&= (\boldsymbol{V}\boldsymbol{\Lambda}^2\boldsymbol{V}^\top)^{-1}\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{U}^\top\boldsymbol{Y} \\
&= \boldsymbol{V}\boldsymbol{\Lambda}^{-2}\boldsymbol{V}^\top\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{U}^\top\boldsymbol{Y} \\
&= \boldsymbol{V}\{(\boldsymbol{\Lambda}^2)^{-1}\boldsymbol{\Lambda}\}\boldsymbol{U}^\top\boldsymbol{Y}. \qquad (1.19)
\end{aligned}$$

In a similar manner, one obtains

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_n^{\text{RR}}(k) &= (\boldsymbol{X}^\top\boldsymbol{X} + k\boldsymbol{I}_p)^{-1}\boldsymbol{X}^\top\boldsymbol{Y} \\
&= (\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^\top + k\boldsymbol{I}_p)^{-1}\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{U}^\top\boldsymbol{Y} \\
&= (\boldsymbol{V}\boldsymbol{\Lambda}^2\boldsymbol{V}^\top + k\boldsymbol{V}\boldsymbol{V}^\top)^{-1}\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{U}^\top\boldsymbol{Y} \\
&= \boldsymbol{V}(\boldsymbol{\Lambda}^2 + k\boldsymbol{I}_p)^{-1}\boldsymbol{V}^\top\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{U}^\top\boldsymbol{Y} \\
&= \boldsymbol{V}\{(\boldsymbol{\Lambda}^2 + k\boldsymbol{I}_p)^{-1}\boldsymbol{\Lambda}\}\boldsymbol{U}^\top\boldsymbol{Y}. \qquad (1.20)
\end{aligned}$$

The difference between the two estimators w.r.t. the SVD is in the bracket terms. Let

$$\boldsymbol{\Lambda}^{-2}\boldsymbol{\Lambda} = \boldsymbol{\Lambda}^{-1} = \text{Diag}\left(\frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_p}\right),$$

$$(\boldsymbol{\Lambda}^2 + k\boldsymbol{I}_p)^{-1}\boldsymbol{\Lambda} = \text{Diag}\left(\frac{\lambda_1}{\lambda_1^2 + k}, \ldots, \frac{\lambda_p}{\lambda_p^2 + k}\right).$$

Since $\lambda_j/(\lambda_j^2 + k) \leq \lambda_j^{-1}$, the ridge penalty shrinks the λ_js. □

1.4 Estimation of Ridge Parameter

We can observe from Eq. (1.18) that the RRE heavily depends on the ridge parameter k. Many authors at different times worked in this area of research and developed and proposed different estimators for k. They considered various models such as linear regression, Poisson regression, and logistic regression models. To mention a few, Hoerl and Kennard (1970), Hoerl et al. (1975), McDonald and Galarneau (1975), Lawless and Wang (1976), Dempster et al. (1977), Gibbons (1981), Kibria (2003), Khalaf and Shukur (2005), Alkhamisi and Shukur (2008), Muniz and Kibria (2009), Gruber et al. (2010), Muniz et al. (2012), Mansson et al. (2010), Hefnawy and Farag (2013), Aslam (2014), and Arashi and Valizadeh (2015), and Kibria and Banik (2016), among others.

1.5 Preliminary Test and Stein-Type Ridge Estimators

In previous sections, we discussed the notion of RRE and how it shrinks the elements of the ordinary LSE. Sometimes, it is needed to shrink the LSE to a subspace defined by $\boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{h}$, where $\boldsymbol{H}$ is a $q \times p$ known matrix of full row rank q ($q \leq p$) and $\boldsymbol{h}$ is a q vector of known constants. It is also termed as constraint or restriction. Such a configuration of the subspace is frequently used in the design of experiments, known as contrasts. Therefore, sometimes shrinking is for two purposes. We refer to this as *double shrinking*.

In general, unlike the Bayesian paradigm, correctness of the prior information $\boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{h}$ can be tested on the basis of samples through testing $\mathcal{H}_{\text{o}} : \boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{h}$ vs. a set of alternatives. Following Fisher's recipe, we use the non-sample information $\boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{h}$; if based on the given sample, we accept $\mathcal{H}_{\text{o}}$. In situations where this prior information is correct, an efficient estimator is the one which satisfies this restriction, called the *restricted estimator*.

To derive the restricted estimator under a multicollinear situation, satisfying the condition $\boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{h}$, one solves the following convex optimization problem,

$$\min_{\boldsymbol{\beta}}\{\text{PS}_{\lambda}(\boldsymbol{\beta})\}, \quad \text{PS}_{\lambda}(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^{\top}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + k\|\boldsymbol{\beta}\|^2 + \lambda^{\top}(\boldsymbol{H}\boldsymbol{\beta} - \boldsymbol{h}), \tag{1.21}$$

where $\lambda = (\lambda_1, \ldots, \lambda_q)^{\top}$ is the vector of Lagrangian multipliers. Grob (2003) proposed the restricted RRE, under a multicollinear situation, by correcting the restricted RRE of Sarkar (1992).

In our case, we consider prior information with the form $\boldsymbol{\beta} = \boldsymbol{0}$, which is a test used for checking goodness of fit. Here, the restricted RRE is simply given by $\hat{\boldsymbol{\beta}}_n(k) = \boldsymbol{0}$, where $\boldsymbol{0}$ is the restricted estimator of $\boldsymbol{\beta}$. Therefore, one uses $\hat{\boldsymbol{\beta}}_n^{\text{RR}}(k)$ if $\mathcal{H}_{\text{o}} : \boldsymbol{\beta} = \boldsymbol{0}$ rejects and $\hat{\boldsymbol{\beta}}_n^{\text{RR(R)}}(k)$ if $\mathcal{H}_{\text{o}} : \boldsymbol{\beta} = \boldsymbol{0}$ accepts. Combining the information existing in both estimators, one may follow an approach by Bancroft (1964), to propose the preliminary test RRE given by

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_n^{\text{RR(PT)}}(k, \alpha) &= \begin{cases} \hat{\boldsymbol{\beta}}_n^{\text{RR}}(k); & \mathcal{H}_{\text{o}} \text{ is rejected} \\ \hat{\boldsymbol{\beta}}_n^{\text{RR(R)}}(k); & \mathcal{H}_{\text{o}} \text{ is accepted} \end{cases} \\
&= \hat{\boldsymbol{\beta}}_n^{\text{RR}}(k) I(\mathcal{H}_{\text{o}} \text{ is rejected}) + \hat{\boldsymbol{\beta}}_n^{\text{RR(R)}}(k) I(\mathcal{H}_{\text{o}} \text{ is accepted}) \\
&= \hat{\boldsymbol{\beta}}_n^{\text{RR}}(k) I(\mathcal{L}_n > F_{p,m}(\alpha)) + \hat{\boldsymbol{\beta}}_n^{\text{RR(R)}}(k) I(\mathcal{L}_n \leq F_{p,m}(\alpha)) \\
&= \hat{\boldsymbol{\beta}}_n^{\text{RR}}(k) - \hat{\boldsymbol{\beta}}_n^{\text{RR}}(k) I(\mathcal{L}_n \leq F_{p,m}(\alpha)),
\end{aligned} \tag{1.22}$$

where $I(A)$ is the indicator function of the set A and $\mathcal{L}_n$ is the test statistic for testing $\mathcal{H}_{\text{o}} : \boldsymbol{\beta} = \boldsymbol{0}$, and $F_{q,m}(\alpha)$ is the upper α-level critical value from the F-distribution with (q, m) degrees of freedom (D.F.) See Judge and Bock (1978) and Saleh (2006) for the test statistic and details.

After some algebra, it can be shown that

$$\hat{\boldsymbol{\beta}}_n^{\mathrm{RR(R)}}(k) = \boldsymbol{R}(k)\hat{\boldsymbol{\beta}}_n, \quad \hat{\boldsymbol{\beta}}_n = \mathbf{0}, \tag{1.23}$$

where

$$\boldsymbol{R}(k) = (\boldsymbol{I}_p + k(\boldsymbol{X}^\top \boldsymbol{X})^{-1})^{-1}.$$

Then, it is easy to show that

$$\hat{\boldsymbol{\beta}}_n^{\mathrm{RR(PT)}}(k,\alpha) = \boldsymbol{R}(k)\hat{\boldsymbol{\beta}}_n^{\mathrm{PT}}, \quad \hat{\boldsymbol{\beta}}_n^{\mathrm{PT}} = \tilde{\boldsymbol{\beta}}_n - \tilde{\boldsymbol{\beta}}_n I(\mathcal{L}_n \le F_{p,m}(\alpha)). \tag{1.24}$$

The preliminary test RRE is discrete in nature and heavily dependent on α, the level of significance. Hence, a continuous α-free estimator is desired. Following James and Stein (1961), the Stein-type RRE is given by

$$\hat{\boldsymbol{\beta}}_n^{\mathrm{RR(S)}}(k) = \boldsymbol{R}(k)\hat{\boldsymbol{\beta}}_n^{\mathrm{S}}; \quad \hat{\boldsymbol{\beta}}_n^{\mathrm{S}} = \tilde{\boldsymbol{\beta}}_n - d\mathcal{L}_n^{-1}\tilde{\boldsymbol{\beta}}_n, \tag{1.25}$$

for some $d > 0$. It is shown in Saleh (2006) that the optimal choice of d is equal to $m(p+2)/p(m+2)$, $m = n - p$.

Note that if $\mathcal{L}_n \to \infty$, then $\hat{\boldsymbol{\beta}}_n^{\mathrm{RR(S)}}(k) = \hat{\boldsymbol{\beta}}_n^{\mathrm{RR}}(k)$, which matches with the estimator $\hat{\boldsymbol{\beta}}_n^{\mathrm{RR(PT)}}(k)$. However, if $\mathcal{L}_n \to 0$, then $\hat{\boldsymbol{\beta}}_n^{\mathrm{RR(S)}}(k)$ becomes a negative estimator. In order to obtain the true value, one has to restrict $(\mathcal{L}_n > d)$. Hence, one may define the positive-rule Stein-type RRE given by

$$\hat{\boldsymbol{\beta}}_n^{\mathrm{RR(S+)}}(k) = \boldsymbol{R}(k)\hat{\boldsymbol{\beta}}_n^{\mathrm{S+}}; \quad \hat{\boldsymbol{\beta}}_n^{\mathrm{S+}} = \hat{\boldsymbol{\beta}}_n^{\mathrm{S}} - (1 - d\mathcal{L}_n^{-1})I(\mathcal{L}_n \le d)\tilde{\boldsymbol{\beta}}_n. \tag{1.26}$$

Although these shrinkage RREs are biased, they outperform the RRE, $\hat{\boldsymbol{\beta}}_n^{\mathrm{RR}}(k)$ in the mean squared error (MSE) sense. One must note that this superiority is not uniform over the parameter space $\mathbb{R}^p$. According to how much $\boldsymbol{\beta}$ deviates from the origin, the superiority changes. We refer to Saleh (2006) for an extensive overview on this topic and statistical properties of the shrinkage estimators. This type of estimator will be studied in more detail and compared with the RRE in the following chapters.

1.6 High-Dimensional Setting

In high-dimensional analysis, the number of variables p is greater than the number of observations, n. In such situations, $\boldsymbol{X}^\top \boldsymbol{X}$ is not invertible and, hence, to derive the LSE of $\boldsymbol{\beta}$, one may use the generalized inverse of $\boldsymbol{X}^\top \boldsymbol{X}$, which does not give a unique solution. However, the RRE can be obtained, regardless of the relationship between p and n. In this section, we briefly mention some related endeavors in manipulating the RRE to adapt with high-dimensional setting. However, in Chapter 11, we discuss the growing dimension, i.e. $p \to \infty$ when the sample size is fixed.

Wang et al. (2016) used this fact and proposed a new approach in high-dimensional regression by considering

$$(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y} = \lim_{k\to 0}(\boldsymbol{X}^\top \boldsymbol{X} + k\boldsymbol{I}_p)^{-1}\boldsymbol{X}^\top \boldsymbol{Y}. \tag{1.27}$$

They used an orthogonal projection of $\boldsymbol{\beta}$ onto the row space of $\boldsymbol{X}$ and proposed the following high-dimensional version of the LSE for dimension reduction:

$$\hat{\boldsymbol{\beta}}_n^{\mathrm{HD}} = \lim_{k\to 0}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top + k\boldsymbol{I}_n)^{-1}\boldsymbol{Y} = \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{Y}. \tag{1.28}$$

They applied the following identity to obtain the estimator:

$$(\boldsymbol{X}^\top \boldsymbol{X} + k\boldsymbol{I}_p)^{-1}\boldsymbol{X}^\top \boldsymbol{Y} = \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top + k\boldsymbol{I}_n)^{-1}\boldsymbol{Y} \tag{1.29}$$

for every $n, p, k > 0$. Buhlmann et al. (2014) also used the projection of $\boldsymbol{\beta}$ onto the row space of $\boldsymbol{X}$ and developed a bias correction in the RRE to propose a bias-corrected RRE for the high-dimensional setting. Shao and Deng (2012) considered the RRE of the projection vector and proposed to threshold the RRE when the projection vector is sparse, in the sense that many of its components are small.

Specifically, write the SVD of $\boldsymbol{X}$ as $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^\top$ where $\boldsymbol{U}(n\times p)$ and $\boldsymbol{V}(p\times p)$ are column orthogonal matrices, and let $\boldsymbol{\Lambda} = \mathrm{Diag}(\lambda_1, \ldots, \lambda_p)$, $\lambda_j > 0$, $j = 1, \ldots, p$ be the eigenvalue matrix corresponding to $\boldsymbol{X}^\top\boldsymbol{X}$. Let

$$\boldsymbol{\theta} = \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{V}\boldsymbol{V}^\top\boldsymbol{\beta}, \quad (\boldsymbol{X}\boldsymbol{X}^\top)^{-} = \boldsymbol{U}\boldsymbol{\Lambda}^{-2}\boldsymbol{U}^\top, \tag{1.30}$$

where $\boldsymbol{\theta}$ is the projection of $\boldsymbol{\beta}$ onto the row space of $\boldsymbol{X}$.

Since $\boldsymbol{X}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-}\boldsymbol{X} = \boldsymbol{X}$, we have $\boldsymbol{X}\boldsymbol{\theta} = \boldsymbol{X}\boldsymbol{\beta}$. Obviously, $\boldsymbol{X}\boldsymbol{\theta} = \boldsymbol{X}\boldsymbol{\beta}$ yields $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$ as the underlying model instead of (1.1). Then, the RRE of the projection is simply

$$\hat{\boldsymbol{\theta}}_n^{\mathrm{RR}}(k_n) = (\boldsymbol{X}^\top \boldsymbol{X} + k_n\boldsymbol{I}_p)^{-1}\boldsymbol{X}^\top \boldsymbol{Y}, \tag{1.31}$$

where $k_n > 0$ is an appropriately chosen regularization parameter.

Shao and Deng (2012) proved that the RRE $\hat{\boldsymbol{\theta}}_n(k_n)$ is consistent for the estimation of any linear combination of $\boldsymbol{\theta}$; however, it is not L_2 consistent, i.e. $n^{-1}\mathbb{E}\|\boldsymbol{X}\hat{\boldsymbol{\theta}}_n(k_n) - \boldsymbol{X}\boldsymbol{\theta}\|^2$ may not converge to zero. Then, they proposed an L_2-consistent RRE by thresholding.

Another approach to deal with high-dimensional data in the case $p > n$ is to partition the vector of regression parameters to main and nuisance effects, where we have p_1 (say) less than n, main and active coefficients that must be estimated and $p_2 = p - p_1$, redundant coefficients that can be eliminated from the analysis by a variable selection method. In fact, we assume some level of sparsity in our inference. Hence, we partition $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)^\top$ in such a way in which $\boldsymbol{\beta}_1$ is a p_1 vector of main effects and $\boldsymbol{\beta}_2$ is a p_2 vector of nuisance effects. We further assume that $p_1 < n$. Accordingly, we have the partition $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ for the design matrix. Then, the multiple linear model (1.1) is rewritten as

$$\boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}. \tag{1.32}$$

This model may be termed as the full model. As soon as a variable selection method is used and p_1 main effects are selected, the sub model has the form

$$\boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}. \tag{1.33}$$

Then, the interest is to estimate the main effects, $\boldsymbol{\beta}_1$ (this technique is discussed in more detail in Chapter 11). Therefore, one needs a variable selection method. Unlike the RRE that only shrinks the coefficients, there are many estimators which simultaneously shrink and select variables in large p and small n problems. The most well-known one is the least absolute shrinkage and selection operator (LASSO) proposed by Tibshirani (1996). He suggested using an absolute-type constraint with form $\sum_{j=1}^{p} |\beta_j| \leq t$ for some positive values t in the minimization of the PRSS rather than the ridge constraint. Specifically, he defined the PRSS by

$$(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + k\|\boldsymbol{\beta}\|^\top\mathbf{1}_p, \quad \|\boldsymbol{\beta}\|^\top\mathbf{1}_p = \sum_{j=1}^{p} |\beta_j|, \tag{1.34}$$

where $\|\boldsymbol{\beta}\| = (|\beta_1|, \ldots, |\beta_p|)^\top$ and $\mathbf{1}_p$ is an n-tuple of 1's. The LASSO estimator, denoted by $\hat{\boldsymbol{\beta}}_n^{\text{LASSO}}$, is the solution to the following convex optimization problem,

$$\min_{\boldsymbol{\beta}} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + k\|\boldsymbol{\beta}\|^\top\mathbf{1}_p. \tag{1.35}$$

To see some recent related endeavors in the context of ridge regression, we refer to Yuzbasi and Ahmed (2015) and Aydin et al. (2016) among others.

In the following section, some of the most recent important references about ridge regression and related topics are listed for more studies.

1.7 Notes and References

The first paper on ridge analysis was by Hoerl (1962); however, the first paper on multicollinearity appeared five years later, roughly speaking, by Farrar and Glauber (1967). Marquardt and Snee (1975) reviewed the theory of ridge regression and its relation to generalized inverse regression. Their study includes several illustrative examples about ridge regression. For the geometry of multicollinearity, see Akdeniz and Ozturk (1981). We also suggest that Gunst (1983) and and Sakallioglu and Akdeniz (1998) not be missed. Gruber (1998) in his monograph motivates the need for using ridge regression and allocated a large portion to the analysis of ridge regression and its generalizations. For historical survey up to 1998, we refer to Gruber (1998).

Beginning from 2000, a comprehensive study in ridge regression is the work of Ozturk and Akdeniz (2000), where the authors provide some solutions for ill-posed inverse problems. Wan (2002) incorporated measure of goodness of fit in evaluating the RRE and proposed a feasible generalized RRE. Kibria

(2003) gave a comprehensive analysis about the estimation of ridge parameter k for the linear regression model. For application of ridge regression in agriculture, see Jamal and Rind (2007). Maronna (2011) proposed an RRE based on repeated M-estimation in robust regression. Saleh et al. (2014) extensively studied the performance of preliminary test and Stein-type ridge estimators in the multivariate-t regression model. Huang et al. (2016) defined a weighted VIF for collinearity diagnostic in generalized linear models.

Arashi et al. (2017) studied the performance of several ridge parameter estimators in a restricted ridge regression model with stochastic constraints. Asar et al. (2017) defined a restricted RRE in the logistic regression model and derived its statistical properties. Roozbeh and Arashi (2016a) developed a new ridge estimator in partial linear models. Roozbeh and Arashi (2016b) used difference methodology to study the performance of an RRE in a partial linear model. Arashi and Valizadeh (2015) compared several estimators for estimating the biasing parameter in the study of partial linear models in the presence of multicollinearity. Roozbeh and Arashi (2013) proposed a feasible RRE in partial linear models and studied its properties in details. Roozbeh et al. (2012) developed RREs in seemingly partial linear models. Recently, Chandrasekhar et al. (2016) proposed the concept of partial ridge regression, which involves selectively adjusting the ridge constants associated with highly collinear variables to control instability in the variances of coefficient estimates. Norouzirad and Arashi (2017) developed shrinkage ridge estimators in the context of robust regression. Fallah et al. (2017) studied the asymptotic performance of a general form of shrinkage ridge estimator. Recently, Norouzirad et al. (2017) proposed improved robust ridge M-estimators and studied their asymptotic behavior.

1.8 Organization of the Book

This book has 12 chapters including this one. In this light we consider the chapter with location and simple linear model first. Chapter 1 presents an introduction to ridge regression and different aspects of it, stressing the multicollinearity problem and its application to high-dimensional problems. Chapter 2 considers simple linear and location models, and provides theoretical developments. Chapters 3 and 4 deal with the analysis of variance (ANOVA) model and seemingly unrelated simple linear models, respectively. Chapter 5 considers ridge regression and LASSO for multiple regression together with preliminary test and Stein-type estimator and a comparison thereof when the design matrix is non-orthogonal. Chapter 6 considers the RRE and its relation to LASSO. Further, we study in detail the properties of the preliminary test and Stein-type estimator with low dimension. In Chapter 7,

we cover the partially linear model and the properties of LASSO, ridge, preliminary test, and the Stein-type estimators. Chapter 8 contains the discussion on the logistic regression model and the related estimators of diverse kinds as described in earlier chapters. Chapter 9 discusses the multiple regression model with autoregressive errors. In Chapter 10, we provide a comparative study of LASSO, ridge, preliminary test, and Stein-type estimators using rank-based theory. In Chapter 11, we discuss the estimation of parameters of a regression model with high dimensions. Finally, we conclude the book with Chapter 12 to illustrate recent applications of ridge, LASSO, and logistic regression to neural networks and big data analysis.

Problems

1.1 Derive the estimates given in (1.12) and (1.13).

1.2 Find the optimum value $k > 0$ for which the MSE of RRE in (1.18) becomes the smallest.

1.3 Derive the bias and MSE functions for the shrinkage RREs given by (1.24)–(1.26).

1.4 Verify that

$$\hat{\boldsymbol{\beta}}_n^{\mathrm{RR}}(k) = (\boldsymbol{I}_p + k(\boldsymbol{X}^\top \boldsymbol{X})^{-1})^{-1} \tilde{\boldsymbol{\beta}}_n.$$

1.5 Verify the identity (1.29).

1.6 Find the solution of the optimization problem (1.35) for the orthogonal case $\boldsymbol{X}^\top \boldsymbol{X} = n\boldsymbol{I}_p$.